Assignment 6. The Great Stanford Hash-Off

Due Friday, March 3 at 1:00 pm

Each student begins with four late days that may be used throughout the quarter. You may submit this assignment 24 hours late by using one late day or 48 hours late using two late days. No submissions will be accepted more than 48 hours after the due date without prior approval by the head TA. See the syllabus for more information about our late policies.

All due dates and submission times are expressed in Pacific time.

You are permitted to work on this assignment in pairs.

Hash tables are one of the most ubiquitous data structures in software engineering, and a lot of effort has been placed into getting them to work quickly and efficiently. In this assignment, you'll explore how efficient those strategies are in practice and will look into their tradeoffs.

This assignment has two main pieces:

- ***Linear Probing:*** You'll implement the linear probing hashing strategy we discussed in class, getting a feel for how to use hash functions and what open addressing looks like in practice.
- ***Performance Analysis:*** Having coded up this hash table, you'll see how fast your code runs on a variety of workflows. From there, you'll explore the performance numbers, identify the tradeoffs between the different hash table representations, and make a recommendation of which hashing strategy you think is best.

As usual, we recommend making slow and steady progress on this assignment throughout the week rather than trying to do everything here in one sitting. Here's our recommended timetable for this assignment:

- Aim to complete Enumerations Warmup the day this assignment goes out.
- Aim to complete Linear Probing Warmup the day this assignment goes out.
- Aim to complete Implementing Linear Probing within five days.
- Aim to complete Performance Analysis component within seven days.

An important note: although the Performance Analysis does not require writing any code, you should make sure to leave appropriate buffer time to complete it. It will require you to think critically about the code you've written and to identify tradeoffs in the implementations. Historically, we've found that students who try doing this part of the assignment the night before it's due get surprised by how much nuanced analysis is required.

Assignment Logistics

Starter Files

We provide a ZIP of the starter project. Download the zip, extract the files, and double-click the `.pro` file to open the project in Qt Creator.

📦 Starter code

Getting Help

Keep an eye on the Ed forum for an announcement of the Assignment 6 **YEAH** (YEAH = Your Early Assignment Help) group session where our veteran section leaders will answer your questions and share pro tips. We know it can be daunting to sit down and break the barrier of starting on a substantial programming assignment – come to YEAH for advice and confidence to get you on your way!

We also here to help if you get run into issues along the way! The Ed forum is open 24/7 for general discussion about the assignment, lecture topics, the C++ language, using Qt, and more. Always start by searching first to see if your question has already been asked and answered before making a new post.

To troubleshoot a problem with your specific code, your best bet is to bring it to the LaIR helper hours or office hours.

Part One: Enumerations Warmup

At a few points in this assignment, you'll be asked to work with ***enumerated types***, a form of custom C++ type representing one out of several different options. We haven't used enumerated types before this assignment, so in this section we figured we'd give you the rundown. 😃

To kick things off, consider the following C++ code:

```
enum class StudentType {
    FROSH,
    SOPHOMORE,
    JUNIOR,
    SENIOR,
    SUPER_SENIOR,
    COTERM,
    MS_STUDENT,
    PHD_STUDENT,
    LAW_STUDENT,
    MED_STUDENT,
    MBA_STUDENT,
    SCPD_STUDENT
};
```

This code creates a new type called **`StudentType`**. This new type is an ***enumerated type***, in which we "enumerate" all the possible values that for a variable of type **`StudentType`**. For example, we can write code like this:

```
StudentType carolyn  = StudentType::FROSH;
StudentType lisa     = StudentType::MS_STUDENT;
StudentType riley    = StudentType::PHD_STUDENT;
StudentType nikolaus = StudentType::LAW_STUDENT;
```

You can compare two **`StudentType`**s against one another using the **`==`** operator, reassign them using **`=`**, etc.

Enumerated types are useful for remembering which of many mutually exclusive options a particular object happens to be in. You'll make use of them in your linear probing hash table to remember whether a slot is empty, full, or a tombstone.

> There are no deliverables for this part of the assignment; just make sure you've read the above code and understand it. 😃

Part Two: Linear Probing Warmup

The linear probing hash strategy that we talked about in Friday's class is very different from the style of hash table (***chained hashing***) that we saw on Wednesday. Here's a few of the differences:

- Each slot in a chained hash table is a bucket that can store any number of elements. Each slot in a linear probing table is either empty or holds a single element.
- Every element in a chained hash table ends up in the slot corresponding to its hash code. Elements in a linear probing table can leak out of their initial slots and end up elsewhere in the table.
- Deletions in a linear probing table use tombstones; deletions in chained hashing don't require tombstones.

Before moving on, we'd like you to answer a few short answer questions to make sure that you're comfortable with linear probing as a collision resolution strategy.

Answer each of the following questions in the file **`ShortAnswers.txt`**.

We have a linear probing table containing ten slots, numbered 0, 1, 2, ..., and 9. For the sake of simplicity, we'll assume that we're hashing integers and that our hash function works by taking the input integer and returning its last digit. (This is a *terrible* hash function, by the way, and no one would actually do this. It's just for the sake of exposition).

Q1. Draw the linear probing table formed by inserting 31, 41, 59, 26, 53, 58, 97, and 93, in that order, into an initially empty table with ten slots. Write out your table by writing out the contents of the slots, in order, marking empty slots with a period (.) character.

Q2. What is the load factor of the table you drew in Q1?

Q3. Draw a *different* linear probing table that could be formed by inserting the same elements given above into an empty, ten-slot table in a different order than the one given above, or tell us that it's not possible to do this. Assume that you're using the same hash function.

Q4. Which slots do you have to look at to see if the table from Q1 – the one formed by inserting the elements in the specific order we gave to you – contains the number 72?

Q5. Which slots do you have to look at to see if the table from Q1 contains the number 137?

Q6. Suppose you remove 41 and 53 from the linear probing table from Q1 using the tombstone deletion strategy described in class. Draw the resulting table, marking each tombstone slot with the letter T.

Q7. Draw the table formed by starting with the table you came up with in Q6 and the inserting the elements 106, 107, and 110, in that order. Don't forget to replace tombstones with newly-inserted values.

Want to check your answers? Stop by the LaIR!

Part Three: Implementing Linear Probing

Your task in this part of the assignment is to implement linear probing in the context of a **`LinearProbingHashTable`** type. Here's the interface for that type, as given in the header file:

```
class LinearProbingHashTable {
public:
    LinearProbingHashTable(HashFunction<std::string> hashFn);
    ~LinearProbingHashTable();

    bool contains(const std::string& key) const;

    bool insert(const std::string& key);
    bool remove(const std::string& key);

    bool isEmpty() const;
    int  size() const;

private:
    enum class SlotType {
        TOMBSTONE, EMPTY, FILLED
    };

    struct Slot {
        std::string value;
        SlotType type;
    };

    Slot* elems;

    /* The rest is up to you to decide; see below */
};
```

Endearing C++ Quirks, Part 1: `string` versus `std::string`

Inside header files, you have to refer to the string type as **`std::string`** rather than just string. Turns out that what we've been calling string is really named **`std::string`**. The line **`using namespace std;`** that you've placed at the top of all your .cpp files essentially says "I'd like to be able to refer to things like **`std::string`**, **`std::cout`**, etc. using their shorter names **`string`** and **`cout`**." The convention in C++ is to not include the using namespace line in header files, so in the header we have to use the full name **`std::string`**.

Think of it like being really polite. Imagine that the string type is a Supreme Court justice, a professor, a doctor, or some other job with a cool honorific, and she happens to be your sister. At home (in your **`.cpp`** file), you call just call her **`string`**, but in public (in the **`.h`** file), you're supposed to refer to her as **`std::string`**, the same way you'd call her Dr. **`string`**, Prof. **`string`**, Justice **`string`**, or whatever other title would be appropriate.

The **`LinearProbingHashTable`** type is analogous to **`Set<string>`** in that it stores a collection of strings with no duplicate elements allowed. However, it differs in a few key ways:

1. The **`LinearProbingHashTable`** type has its hash function provided to it when it's created. In practice, a hash table should choose a hash function internally. We have set up the **`LinearProbingHashTable`** type this way for the purposes of this assignment to make testing easier; we can construct tests where we specify the hash function and know exactly where each element is going to land.

2. The **`Set<string>`** type has no upper bound to how big it can get. For reasons that we'll discuss later, the **`LinearProbingHashTable`** type you'll be implementing is built to have a fixed number of slots, which is specified in the constructor. (Specifically, you're given a particular **`HashFunction<string>`**, and that **`HashFunction<string>`** is

built to work with a specific number of slots.) This means that there's a maximum number of elements that can be stored in the table, since a linear probing table can't store more than one element per slot.

In terms of the internal representation of the **`LinearProbingHashTable`**: as with the **`HeapPQueue`**, you also need to do all your own memory management, though we suspect this will be easier than the **`HeapPQueue`** because the size of your hash table never changes. You should represent the hash table as an array of objects of the type **`Slot`**, where **`Slot`** is the **`struct`** type defined inside **`LinearProbingHashTable`**. Each slot consists of a string, along with a variable of the enumerated type **`SlotType`** indicating whether the slot is empty, full, or a tombstone. If the slot is empty or a tombstone, you should completely ignore the string value, since we're pretending the slot is empty in that case. If the slot is not empty, the string value tells you what's stored in that particular slot.

We've provided you with a fairly extensive set of automated tests you can use to validate that your implementation works correctly. To assist you with testing, we've also provided an "Interactive Linear Probing" environment akin to Assignment 5's "Interactive PQueue" button that lets you issue individual commands to a linear probing table to see what happens.

Linear Probing Requirements

Here's our recommendation for how to complete this assignment:

Implement the **`LinearProbingHashTable`** type in **`LinearProbingHashTable.h/.cpp`**. To do so:

1. Read over **`LinearProbingHashTable.h`** to make sure you understand what all the functions you'll be writing are supposed to do.

2. Add some member variables to **`LinearProbingHashTable.h`** so that you can, at a bare minimum, remember the hash function given to you in the constructor. (You'll may or many not need more member variables later; we're going to leave that up to you.) Remember that you need to do all your own memory management.

3. Implement the constructor, which should create a table filled with empty slots and store the hash function for later use. You can determine how many slots your table should have by calling the **`HashFunction<T>::numSlots()`** member function on **`hashFn`**, which returns the number of slots that the hash function was constructed to work with.

4. Implement the **`size()`** and **`isEmpty()`** functions, along with the destructor. Both functions should run in time O(1). The **`size()`** member function should return the number of elements currently in the table, rather than the number of slots in the table. (Do you see the distinction?)

5. Implement **`contains()`** and **`insert()`**. For now, don't worry about tombstones or removing elements. You should aim to get the basic linear probing algorithm working correctly.

6. Confirm that you pass all the tests that don't involve removing elements from the table, including the stress tests, which should take at most a couple of seconds each to complete. Don't forget that, if you aren't passing a test, you can set a breakpoint in the test and then run your code in the debugger to step through what's going on. You can also use the Interactive Linear Probing option to explore your table in an interactive environment – preferably with the debugger engaged.

7. Implement the **`remove()`** function. You should use the tombstone deletion algorithm described in lecture. This may require you to change the code you've written so far:

    1. You may – or may not – need to change your code for **`contains()`** to handle tombstones. It depends on how you implemented **`contains()`**.

2. You may – or may not – need to change your code for **`insert()`** to handle tombstones. Remember to place elements in the first empty or tombstone slot that you find. (And make sure not to insert an element into the table if it's already there!)

8. Confirm that you pass all the provided tests, including the stress tests.

Some notes on this problem:

- ***Do not implement contains, remove, or insert recursively.*** Some of the stress tests we'll be subjecting your code to in the time tests will involve working with extremely full tables, and you can easily get a stack overflow this way because the call stack won't be big enough to hold space for all the stack frames.

- Make sure you store the hash function that we provide you in the constructor and use that hash function throughout the table. Variables of type **`HashFunction<string>`** are like other types of variables; you can assign them by writing code to the effect of **`hashFn1 = hashFn2;`**. In the event that you have a **`HashFunction<string>`** as a data member (e.g. something in the **`private`** section of the class) that has the same name as a local variable or parameter to a function, write **`this->hashFn = hashFn;`**.

- If you have a variable of type **`HashFunction<string>`** named **`hashFn`** and an element to hash named **`elem`**, you can compute the hash code of **`elem`** by writing **`hashFn(elem)`**.

- Computing the hash code of a string is fast but not instantaneous. To avoid introducing inefficiencies into your code, try to avoid recomputing the same item's hash code multiple times in a loop when performing an insertion, deletion, etc.

- Remember that, like the **`Set<string>`** type, your **`LinearProbingHashTable`** should not allow for duplicate elements. If the user wants to insert a string that's already present, you should not insert a second copy.

- In Assignment 5, we added some code to the **`DataPoint`** type to make it a lot easier to spot when you had accidentally walked off the end of an array or otherwise used an uninitialized **`DataPoint`**. We have not provided the same safeguards here. This means that if you walk off the end of your array of slots, or try reading from a negative index, the behavior you see may vary from run to run. Your code might work correctly sometimes, crash other times, or produce test failures in other runs.

- A common mistake in implementing this hash table is to try to set the **`value`** field of a **`Slot`** to **`nullptr`** when removing a string from the hash table, with the idea of saying "this string doesn't exist any more." That, unfortunately, doesn't work. Remember that in C++, a string represents an honest-to-goodness string object, so you can't have a "null string" the same way you can't have a "null integer." Unfortunately, it is legal C++ code to assign **`nullptr`** to a string, which C++ interprets as "please crash my program as soon as I get here" rather than "make a null string." Instead, just change the type field to indicate that although there is technically a string in the slot, that string isn't meaningful.

- Make sure not to read the contents of a string in a **`Slot`** if that **`Slot`** is empty or a tombstone. If the slot isn't filled, it means that the string value there isn't meaningful. There are a lot of bugs that can arise if you accidentally read the string in a **`Slot`** when the slot doesn't have an element in it.

- Your table should be able to store any strings that the user wants to store, including the empty string.

- The **`contains`**, **`insert`**, and **`remove`** functions need to function correctly even if the table is full. Specifically, **`insert`** should return false because there is no more space, and **`remove`** and **`contains`** should operate as usual. You may need to special-case the logic here – do you see why?

- You are encouraged to add private helper functions, especially if you find yourself writing the same code over and over again. Just don't change the signatures of any of the existing functions. If you do define any helper functions, think about whether they

should be marked **const**. Specifically, helper functions that don't change the hash table should be marked **const**, while helper functions that do make changes to the table should not be **const**.

- You are likely to run into some interesting bugs in the course of coding this one up, and when you do, don't forget how powerful a tool the debugger is! Feel free to set breakpoints in the different test cases so that you can see exactly what your code is doing when it works and when it doesn't work. Inspect the contents of your array of slots and make sure that it's consistent with what you expect to see. Once you've identified the bug – and no sooner – edit your code to fix the underlying problem.

- You ***must not*** use any of the container types (e.g. **Vector**, **Set**, etc.) when solving this problem. Part of the purpose of this assignment is to let you see how you'd build all the containers up from scratch.

- We encourage you to write your own **STUDENT_TEST**s here to help debug your solution as you go. However, you are not required to do so.

Endearing C++ Quirks, Part 2: Returning Nested Types

There's another charming personality trait of C++ that pops up when implementing member functions that return nested types. For example, suppose that you want to write a helper function in your **LinearProbingHashTable** that returns a pointer to a **Slot**, like this:

```
private:
    Slot* hsAreCute();
```

In the **.cpp** file, when you're implementing this function, you need to give the full name of the **Slot** type when specifying the return type:

```
LinearProbingHashTable::Slot* LinearProbingHashTable::hsAreCute() {
    // Wow, this pun lost a lot in translation.
}
```

While you need to use the full name **LinearProbingHashTable::Slot** in the return type of an implementation of a helper function, you don't need to do this anywhere else. For example, this code is perfectly legal:

```
LinearProbingHashTable::Slot* LinearProbingHashTable::hsAreCute() {
    Slot* h = new Slot[137]; // Totally fine!
    return h;
}
```

Similarly, you don't need to do this if the function takes a **Slot** as a parameter. For example, imagine you have this member function:

```
private:
    void iLostMoneyToA(Slot* machine);
```

You could implement this function without issue as

```
void LinearProbingHashTable::iLostMoneyToA(Slot* machine) {
    // Don't make the same mistake as me!
}
```

Part Four: Performance Analysis

We implemented chained hashing in class, and we've included a **`ChainedHashTable`** type in the **`Demos/`** directory along the lines of the hash table you built here. You just implemented a linear probing table. The question then is – how do they stack up against one another?

Choose the "Performance Analysis" button from the main menu. This option will run the following workflow on each of the table types:

- Insert all words in the file **`EnglishWords.txt`** into an empty hash table. Each item is inserted twice, the first time to measure the speed of a successful insertion, and the second time to measure the speed of an unsuccessful insertion when the item is already present.
- Look up each word in **`EnglishWords.txt`** in that hash table, measuring the average cost of each successful lookup.
- Look up the capitalized version of each word in that hash table. Since all the words in **`EnglishWords.txt`** are stored in lower-case, this measures the average cost of each unsuccessful lookup.
- Remove the capitalized versions of each word in the hash table. This measures the average cost of unsuccessful deletions, since none of those words are present.
- Remove the lower-case versions of each word in the hash table. This measures the average cost of successful deletions.

The provided starter code will run this workflow across a variety of different load factors α (ratios of numbers of elements in the table to the number of slots in the table), reporting the times back to you. ***This may take a while to complete;*** it's okay if it takes about five or ten minutes to finish running.

As a note, ***the timing numbers you get will be sensitive to what else is running on your computer.*** If you leave your program running time tests in the background while, say, watching a YouTube video, the overhead of your computer switching back and forth between different processes can skew the numbers you'll get back. We recommend that once you click the "Performance Analysis" button, you walk away from your computer for a while, stretch a bit, and return once all the time trials have finished.

Once you have the data, review the numbers that you're seeing. Look vertically to see how the times for a particular hash table compare across load factors, and horizontally to see how the different tables compare against one another.

Answer each of the following questions in the file **`HashTableAnalysis.txt`**.

Q1. Look at the costs of ***successful lookups*** in a linear probing hashing across a range of load factors. Describe any patterns that you see. Then, based on how the linear probing lookup algorithm works, explain why you're seeing what you're seeing. (Your answer should be 2 - 4 sentences in length.)

Q2. Repeat the above exercise for ***unsuccessful lookups***.

Q3. Repeat the above exercise for ***successful insertions***.

Q4. Repeat the above exercise for ***unsuccessful insertions***.

Q5. Repeat the above exercise for ***successful deletions***.

Q6. Repeat the above exercise for ***unsuccessful deletions***.

Q7. Explain why it would *not* be a good idea to use a linear probing hash table with a load factor of 0.01 even though it would be much faster than with a higher load factor.

Q8. In practice, someone has to make a judgment call about which type of hash table to use and with which load factors. Review the numbers you've seen for chained hashing and linear probing hashing. Propose a single choice of hash table and load

factor that you believe is the "best," then justify your decision. (Your answer should be 2 - 4 sentences in length.)

(Optional) Part Five: Extensions

Want to go above and beyond? Here's some suggestions of how to get started.

- We described Robin Hood hashing in class, but didn't ask you to code it up here. If you'd like to take a stab at coding that one up, implement Robin Hood hashing in the files **`Extras/RobinHoodHashTable.h`** and **`Extras/RobinHoodHashTable.cpp`**. You can then modify the file **`Demos/TimeTestConfig.h`** to hook your Robin Hood hash table into our time tests. How does Robin Hood hashing compare to linear probing and to chained hashing?

- If that isn't enough for you, there are many other hashing strategies you can use. Quadratic probing and double hashing, for example, are variations on linear probing that cap the number of elements that can be in a slot at one, but choose a different set of follow-up slots to then look at when finding the next place to look. Cuckoo hashing is based on a totally different idea: it uses two separate hash functions and places each element into one of two tables, always ensuring that the elements are either at the spot in the first table given by the first hash function or the spot in the second table given by the second hash function. Hopscotch hashing is like linear probing, but ensures that elements are never “too far” away from their home location. FKS hashing is like chained hashing, but uses a two-layer hashing scheme to ensure that each element can be found in at most two probes. Read up on one of these strategies – or another of your choosing – and code them up. How quickly do they run? You can add your own hash table type to the performance analysis by editing **`Demos/TimeTestConfig.h`**.

- Alternatively, you could explore ways of making your linear probing table more "realistic." The hash tables we’ve defined here are fixed-sized and don’t grow when they start to fill up. In practice, you’d pick some load factor and rehash the tables whenever that load factor was reached. Write code that lets you rehash the tables once they exceed some load factor of your choosing. Tinker around to see what the optimal load factor appears to be! Alternatively, look into the following. Tombstone deletion has a major drawback: if you fill a linear probing table up, then delete most of its elements, the cost of doing a lookup will be way higher than if you had just built a brand new table and filled it in with just those elements. (Do you see why?) Some implementations of these hash tables will keep track of how many deleted elements there are, and when that exceeds some threshold, they’ll rebuild the table from scratch to clean out the tombstones. Experiment with this and see what you can come up with!

- We provide the hash functions in this assignment, but there’s no reason to suspect that our hash functions are the “best” hash functions and you can change which hash function to use. Research other hash functions, code them up, and update the performance test (it’s the **`timeTest`** function in the files **`Demos/PerformanceGUI.cpp`**) to use your new hash function. How does your new hash function compare with ours?

Submission Instructions

Before you call it done, run through our submit checklist to be sure all your **`t`**s are crossed and **`i`**s are dotted. Make sure your code follows our style guide. Then upload your completed files to Paperless for grading.

Partner Submissions:

- If you forget to list your partner you can resubmit to add one
- Either person can list the other, and the submissions (both past and future) will be combined
- Partners are listed per-assignment

- You can't change/remove a partner on an individual submission

Please submit only the files you edited; for this assignment, these files will be:

- `ShortAnswers.txt`. ***(Don't forget this one, even though there's no code in it!)***
- `HashTableAnalysis.txt`. ***(Don't forget this one, even though there's no code in it!)***
- `LinearProbingHashTable.h/.cpp`. ***(Remember to submit both of these files!)***

You don't need to submit any of the other files in the project folder.

Submit to Paperless

If you modified any other files that you modified in the course of coding up your solutions, submit those as well. And that's it! You're done!

Good luck, and have fun!